Knowledge management by information mining

Özer Ciftcioglu and Sanja Durmisevic
Delft University of Technology

Key words: Knowledge management, Information mining, Sensitivity analysis

Abstract: Novel information mining method dealing with soft computing is described. By this method, in the first step, receptive fields of design information are identified so that connections among various design aspects are structured. By means of this, complex relationships among various design aspects are modeled with a paradigm, which is non-parametric and generic. In the second step, the structured connections between various pairs of aspects are graded according to the relevancy to each other. This is accomplished by means of *sensitivity analysis,* which is a computational tool operating on the model established and based on a concept measuring the degree of dependencies between pairs of quantities. The degree of relationships among various design aspects so determined enables one to select the most important independent aspects in the context of design or decision-making process. The paper deals with the description of the method and presents an architectural case study where numerical and as well as non-numerical (linguistic) design information are treated together, demonstrating a ranked or elective information employment which can be of great value for possible design intervention during reconstruction.

1. INTRODUCTION

In architectural design process, one has to establish certain relations among the design information in advance to make design with sound rationales behind. The main difficulty at this point is that such relationships may not be determined because of various reasons. One example may be the vagueness of the architectural design data due to linguistic qualities in them. Another example may be the vaguely defined design qualities, which should be gradually fixed during the actual implementation, in order to maintain the flexibility of the design for architectural real-time decision-makings. To deal

with such flexible design information is not an easy task since the majority of the existing architectural design aids, so-called decision support systems, are based on concrete input design information to be provided and well-defined final goals. Here the problem is not only the initial fuzziness of the information but also the desired relevancy determination among all pieces of information given. Basically, if there is no relevancy between two aspects, they need not to be considered in that context of design. Acquiring this knowledge in advance might save considerable design efforts in the meanwhile. However presently, to determine the existence of such a relevancy is more or less a matter of architectural subjective judgement rather than a systematic non-subjective decision-making based on the existing design information. In this respect, the invocation of certain design tools dealing with such fuzzy information is essential for enhanced design decisions. Referring to the fuzziness of the information subject to treatment, naturally fuzzy logic tools are most appropriate to employ (Tanaka, 1997).

2. KNOWLEDGE MANAGEMENT

Knowledge is the formation of implicit and explicit restrictions placed upon objects, operations and relationships along with general and specific heuristics and inference procedures involved in the situation being modeled (Sowa, 1985). From the design activity viewpoint, *specific heuristics and inference procedures* play important role since there is no firm guidelines for a specific design task. Here, heuristic may be understood as any effort based on the performance of the case and intended for the accomplishment of the related task. In a design task, specific heuristic and inference procedures may play important role. This is especially the case if there are essential imprecision in the data and this is the case in this research. The imprecision in the data can be handled by fuzzy logic techniques so that generic information about the relevancy among the design items and the effectiveness of each design item in this context can be identified. This is a basic information mining procedure for the elicitation of information from the data. Information mining and knowledge management is essential activity in decision support systems with inference capability. Fuzzy expert systems can provide the required flexibility for dealing with the imprecise data and basically such a scheme is described in *figure 1*.

In a complex information environment, to establish the complete fuzzy rules dealing with the knowledge base is a formidable task. To alleviate the problem, the knowledge base can be formed in a distributed and structured form by means of learning (machine learning) so that the structure represents the fuzzy expert system altogether with consistent rules in any complexity.

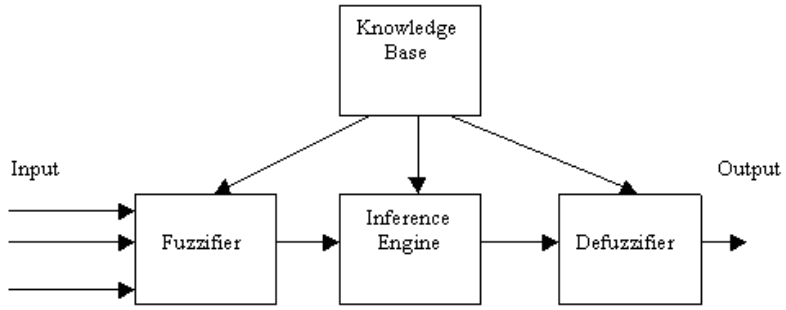

Figure 1. Scheme of fuzzy expert system

Here the main task is the establishment of the consistent rules. This can be achieved by performing information-mining methods with appropriate learning process, which should be specially designed for this purpose. Information mining is closely related to data mining. The main difference between information mining and conventional data reduction is the complexity of the data and the method of elicitation of the information being searched for. Data mining is in essence is the process of data reduction (Han, Kamber 2001). It is one of the dominant techniques of exploratory data analysis and there are a number of ways for it. Clustering is one of the essential methodologies of concern (Han, Kamber 2001; Krzysztof, Pedrycz, et al. 1998). It can broadly be divided into two categories as clustering with unsupervised and clustering with supervised learning. In the unsupervised case, given an amount of information, data set and the distance function, one starts with the number of clusters assigning each pattern to a separate cluster and proceeds to merge the clusters that are the closest. In the supervised case, we are concerned with a collection of labeled data that come in the form of ordered pairs. This can be seen as a feature vector describing the data and its class assignment. Eventually, the associations between the feature vectors and the class assignments are endeavored for determinations. Perhaps the major disadvantage of unsupervised clustering is due to fact that the clusters are established from the input data and the final associations with some output functions are performed afterwards. In this case the associations may not be optimal since the established clusters may not be the optimal representations with respect to associations being looked for. In the supervised case the clustering and associations are performed in a constructive way so that the most effective representation of associations being investigated is obtained. The accomplishment of this structure can be realized by means of a network, which has special knowledge management features suitable for dealing with imprecise data. Such a structure can be a radial basis function network (RBF) network. RBF networks form one of the

essential categories of neural networks. The main architectures, learning abilities, and applications are described in the literature (Brommhead, Lowe 1988; Moody, Darken 1989; Park, Sandberg 1991; Leonard, Kramer, et al. 1992; Chen, Manry 1993; Eleneyar, Shin 1994) where the learning and generalization abilities of these networks are outstanding. In particular, some interesting equivalence between RBF networks and fuzzy rule-based systems have been established (Jang, Sun 1993; Hunt, Haas, et al. 1998). Here, the antecedants (input premises) of the respective rules describing the fuzzy model give rise to a linguistic partition of the input space. An RBF network operates with fuzzy computational units, which are in essence cluster centers and they constitute the essential functional components in the structure. However, from information mining and knowledge management viewpoint, these cluster centers are called receptive fields to distinguish between information mining and data mining since selection of the receptive fields differs from conventional clustering techniques as explained below.

Referring to the form of the computational units in an RBF structure, especially their distribution in the input space is often critical for the performance of the network. In general, in the knowledge model by RBF networks, categorically two stages can be identified. In the first stage, encapsulation of some domain knowledge is carried out. In the second stage, a parametric learning process takes place. For the encapsulation of the domain knowledge there are two main sources of this knowledge:

- *A preliminary analysis of training data* where the data set is preliminarily analyzed and the receptive fields are determined. For this, general data clustering methods are of particular interest (Bezdek 1981).
- *Designer-oriented data analysis* where receptive fields are formed based upon intended design goals of designer. In this respect, the preceding category concerns self-organization and this category belongs to supervised organization. In this second stage, the domain knowledge can be more explicitly emphasized in the model. This allows focusing attention on the machine learning on some essential regions of the input space. As a result of this a multiresolutional character in knowledge modeling together with efficient learning is exercised.

Concerning information mining and knowledge management by RBF networks information is embedded locally in the form of a database where elicitation of information from the data requires special methods and techniques. In particular, referring to the encapsulation of the knowledge, the receptive fields are formed by means of second categorical source of information described above, where special type of supervised clustering for determination of the receptive fields is performed, as a first step. A parametric learning follows this, as a second step. These two steps in knowledge modeling are imperatively performed by orthogonal least squares

(OLS) algorithm (Chen, Cowan, et al., 1991). In particular, by conventional clustering, cluster centers are located anywhere in the input space, matching the clustering process to some prescribed clustering criteria. This is purely a mathematical treatment and the centers identified might have no correspondence to a physical entity or reality. However, for knowledge management, the model for knowledge base should be constructed on actual data rather than some mathematical abstractions derived from data. In this respect, in the OLS, the centers are part of the data as receptive fields rather than clusters. This means, each receptive field is a set of data in the form of a data vector in a multidimensional input space. More explicitly, they are a part of the data reflecting the exact information present in the data to the knowledge model without any mathematical abstraction. They refer to the appropriate central locations in the model according to data at hand. By doing so, the domain knowledge is effectively emphasized by means of two important gains accomplished. In the first place, the effective management of knowledge by the appropriate distribution of the receptive fields is carried out. In the second place, the enhanced generalization capability of the knowledge model is guaranteed by selecting appropriate number of receptive fields in the model. In general, the selection of number of clusters or receptive fields in such a model is critical and therefore its appropriate selection is essential concern. In the literature, this phenomenon is referred to as *bias-variance dilemma* (Duda, Peter, et al., 2001) and mostly treated in the context of neural networks (Haykin, 2000). This phenomenon plays also important role on the model developed in this research. The dilemma basically manifests itself by the estimations through the model. If the number of receptive fields in the model are excessive, than the model errors with the training data are relatively small while the same errors are relatively high for unknown data applied to the input of the model, and vice versa. Since overall effect on the quality of the knowledge model is the combination of these two conflicting errors, the model should have appropriate number of receptive fields as optimal (see *figure 2*). The model error obtained from the knowledge model with unknown data at its input is used as a measure for performance of the model and it is referred to as *generalization error.* In this context, the performance of the model is expressed in terms of its *generalization capability* measured by the generalization error.

3. KNOWLEDGE MANAGEMENT FOR ARCHITECTURAL DESIGN - CASE STUDY

In this research, knowledge management by information mining is performed for an architectural study related to underground station design. The data collected comprised various design features, in general. The knowledge management considers certain aspects of the station design. For this purpose the dependency among various design features is required. This dependency can be conceived as an input and output structure where the inputs are mapped to the outputs via a knowledge model. Therefore, from the data at hand the general dependency information was needed to devise a model in a structural and continuous form. Such a model could be used for any set of design requirements as a knowledge base where for these requirements at the input, it provides consistent design solutions at the output.

In basic mathematical terms, if we assume that the input vector x(t) represents a n-dimensional vector of real-valued fuzzy membership grades: $x(t) \in [0,1]^n$, then, the output y(t) of the model represents an m-dimensional vector of corresponding real-valued membership grades, $y(t) \in [0,1]^m$. This structure actually performs a nonlinear mapping from an n-dimensional hypercube $I^n=[0,1]^n$ to an m-dimensional hypercube $I^m=[0,1]^m$:

$$x(t) \in [0,1]^n \rightarrow y(t) \in [0,1]^m$$

In the actual implementation, $x_i(t)$ (i=1,2,…,n) in x(t) represents the degree to which an input fuzzy variable $x_i'(t)$ belongs to a fuzzy set, while $y_j(t)$ (j=1,2,…,m) in y(t) represents a degree to which an output fuzzy variable $y_j'(t)$ belongs to a fuzzy set. Further in the text, a case study is presented and based on that data sensitivity analysis is conducted.

A data set obtained for Blaak underground station in the Netherlands was used for training as a case study. This is an important exchange station, which is situated in the center of Rotterdam. It is at the same time a tram, metro and a train station. Tram station is situated on a ground level. Metro platforms are one level below ground (at approximately -7m) and train platforms are two levels below ground (at approximately -14 meters). From 27th May till 30th May 2000, one thousand of questionnaires were handed out to the passengers visiting the station. The questionnaire covered aspects that are related to safety and comfort at the station. In total there were 43 aspects in input space each having five possible options and two aspects in the output space again having five possible options. The latter two are design variables being *safety* and *comfort*. The input aspects, which are identified to be related to comfort are given in *table 1* and those related to safety are

presented in *table 2* (Ciftcioglu, Durmisevic, et al., 2001). Main purpose of the questionnaire was to provide information on user's perception regarding specific spatial characteristics of that station. The questions covered all aspects given in *table 1* and *table 2* and additional two final questions were related to users' perception of public safety and comfort at Blaak station.

In general, the way in which data is measured is called a level of measurement or the scale of measurement for variables. The organization of variables determines how data can be analyzed (McGrew, Monroe, 1993). There are four kinds of scales in which variables can be measured: the *nominal* scale, the *ordinal* scale, the *interval* scale and the *ratio* scale (Dalen, Leede, 2000). In this research we have used an *ordinal scale measurement*, where such measurement involves placement of values in a rank order, in order to create an ordinal scale variable. The relationship between observations takes on a form of 'greater than 'and 'less than'. In the questionnaire, variables are divided in a five-point scale. The use of a three-point scale or a five-point scale in a questionnaire are the most common (Baarda , de Goede, 1997). For this questionnaire a five-point scale was used in order to generate a distinct approach. In such way, respondents have the opportunity to 'strongly agree' or 'strongly disagree' with a question or to strongly express their opinion regarding design issues.

Table 1. Aspects related to comfort (total 28)

Attractiveness	**Wayfinding**	**Daylight**	**Physiological**
Colour	To the station	Pleasantness	Noise
Material	In station	Orientation	Temperature winter
Spatial proportions	Placement signs		Temperature summer
Furniture	Number of signs		Draft entrance
Maintenance			Draft platforms
Spaciousness entrance			Draft exchange areas
Spaciousness train platform			Ventilation entrance
Spaciousness metro platform			Ventilation platforms
Platform length			
Platform width			
Platform height			
Pleasantness entrance			
Pleasantness train platform			
Pleasantness metro platform			

Table 2. Aspects related to safety (total 15)

Overview	**Escape**	**Lighting**	**Presence of people**	**Safety surr.**
Entrance	Possibilities	Entrance	Public control	Safety in surrounding
Train platform	Distances	Train platform	Few people daytime	
Metro platform		Metro platform	Few people night	
Exchange area		Exchange area		
		Dark areas		

In a 6 weeks period 219 completed questionnaires were returned. Some cases were immediately excluded since they failed the control question, which was in the questionnaire in order to improve reliability of the outcomes. For training we used in total 196 cases and 7 cases were used in order to test the network performance.

Figure 2 provides the outcomes of the network training. *Figure 3* provides validation of the network performance for both comfort and safety aspects. For training 196 cases were used and for the knowledge model validation, the number of receptive fields selected was 90 out of graded sequence of centers which was in total 196. The dimension of the input space was 43. The network estimation is rather reasonable considering the fact that there are a number of possible combinations in the input space that can be represented as input information. More precisely at the order of 5^{43} different combinations are possible, since each dimension has five options, referring to earlier mentioned questionnaire. Thanks to the information mining in the form of receptive field clustering, in such a large dimensional input space, relatively small number of convergence points (altogether 90 receptive fields out of 196 input sets of data) are precisely located. It is important to point out that, the number of receptive fields is found to be optimal around 90 for this particular model and this is related to the bias variance dilemma mentioned before.

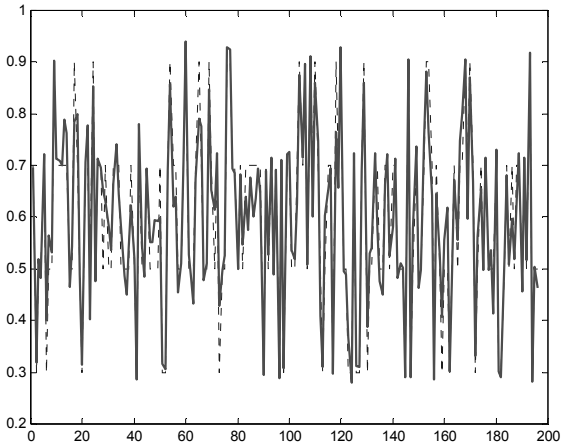

Figure 2. Network training for 196 cases with 43 inputs and 1 output as *comfort* variable with 90 receptive fields. Broken lines represent the knowledge model response to the training data after training. Continuous lines represent the actual knowledge used for modeling. Some difference is visible and referring to the bias-variance dilemma, it has positive effect for the generalization capability of the model (see Figure 3)

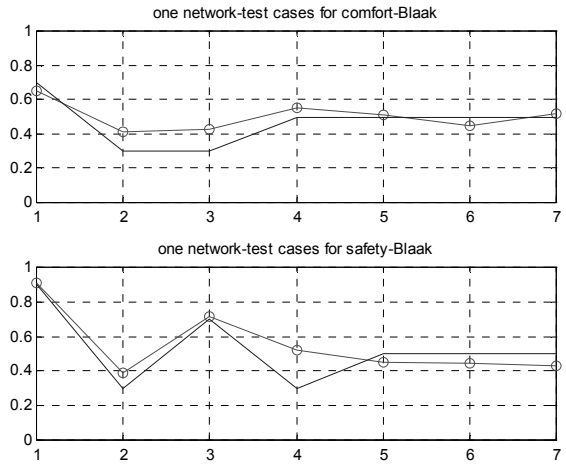

Figure 3. Test results of the network performance for comfort (upper picture) and safety (lower picture). Line with circles represents the estimated value by the network, while the other line is the actual value

In *figure 2*, the broken lines represent the knowledge model response to the training data after training. The difference between the actual data and

model response is basically modeling error, which is deliberately allowed to enhance the generalization capability of the model.

It is interesting to note that in *figure 2*, for the number of receptive fields equal to the number of input sets of data, which is 196 in this case, the modeling error virtually vanishes at the price of large generalization errors.

Having trained the network, whole information is stored in a compact way, which makes further knowledge management more efficient. It is also important to mention that it is necessary to train the network including aspects for comfort and safety at the same time, rather than to separate and train as two independent networks. The reason for that is rather obvious, since comfort and safety are very much related to each other, and certain issues related to one are embedded in the other as well and therefore would be rather superficial to separate them as independent information in the network training. This was confirmed through experiments as well (Durmisevic, Ciftcioglu, et al., 2001).

The relative dependency of the input variables on comfort and safety is identified by means of sensitivity analysis (Saltelli, Chan, et al. 2000) where basically the gradients of comfort and safety with respect to each variable in the input space is computed. The results are shown in *figure 4*, where the '*x*' axis represents 43 dimensions in the input space, and the '*y*' axis is a sensitivity of that dimension to the comfort and safety at the output, ranked on a scale from 0 to 1. From this figure, In *table 3* and *table 4*, only some most sensitive aspects are mentioned as indication of sensitivity analysis results. Since the computation is possible with analytical computation accurate results are obtained.

Table 3. Order of aspects that are most sensitive to feeling of comfort at Blaak station

Number	relative importance value	Sensitivity to comfort
1	1.0000	Spaciousness metro platform
2	0.8972	Spatial proportions
3	0.8723	Platform width
4	0.8171	Pleasantness metro platform
5	0.7647	Platform height
6	0.7403	Spaciousness entrance
7	0.6738	Pleasantness train platform
8	0.6508	Lighting of train platform
9	0.6102	Platform length

Table 4. Order of aspects that are most sensitive to feeling of safety at Blaak station

Number	relative importance value	sensitivity to safety
1	1.0000	Safety in surrounding
2	0.8202	Few people present during night
3	0.7266	Few people present during daytime
4	0.3680	Pleasantness of metro platform

Number	relative importance value	sensitivity to safety
5	0.3517	Lighting of train platform
6	0.3239	Wayfinding in station

On one hand, since the first 28 variables are directly related to comfort, the associations of these variables to comfort are prevailing in that region (*see table 1 and figure 4 upper picture for comparison*). On the other hand, since the last 15 variables are directly related to safety, the associations of these variables to safety are clearly prevailing in that region (*see table 2 and figure 4 lower picture for comparison*). Yet it is noticeable that certain aspects for safety are very much influencing comfort and other way round. One example is an aspect 36 being 'lighting at the train platform'. When we compare two sensitivity analysis results for comfort and safety we notice that this aspect is very sensitive to comfort, even though in the first place it was assumed that it might be more related to safety (*see table 2*). In that sense, sensitivity analysis proved to be an effective method indicating validity/invalidity of assumptions made. This provides additional confidence on the quality of the knowledge management and the validity of the knowledge model developed.

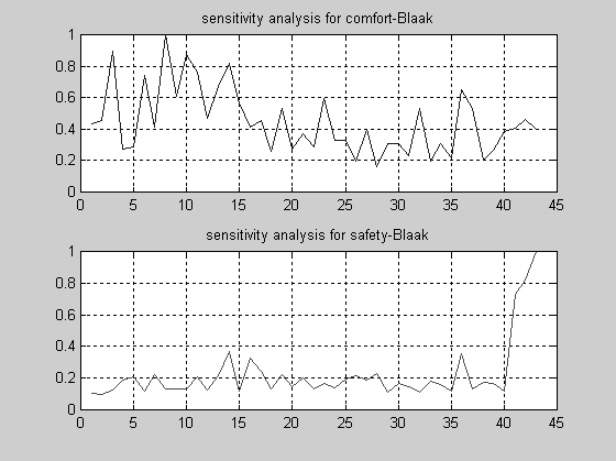

Figure 4. Upper figure is the sensitivity of comfort to input space variables and lower figure is the sensitivity of safety to the input space variables

The following table (*table 5*) gives an insight into final outcomes of estimation of comfort and safety provided by the passengers at Blaak station.

Table 5. [Evaluation of comfort and safety aspects for the Blaak station expressed in %]

5 scale measurement	Comfort	Safety	5 scale measurement
Very uncomfortable	1.4	3.7	Very unsafe
Uncomfortable	15.9	16.4	Unsafe
Reasonably comfortable	35.5	35.5	Reasonably safe
Comfortable	36.4	30.8	Safe
Very comfortable	10.7	13.6	Very safe

The outcomes from the knowledge model indicated outstanding potentiality of the model for gaining detailed insight into the information at hand and the effective use of such design information in a structural form as a knowledge base, for enhanced architectural design decisions. In the context of knowledge management, this knowledge model was especially designed to assess the qualitative aspects of design, including technical and quantitative values. The results of this study are valuable for eventual reconstruction of Blaak station, since it provides an architect with a valuable information regarding different aspects considered in this research. Numerous conclusions can be derived out of this model, like for example, the fact that the safety in the surrounding is the most determining aspect for feeling safe in the station. This means that no matter which architectural intervention are made inside the station, if a designer does not consider the surrounding where this station is situated and possible improvements in that surrounding, it is less likely that people will feel safer at that station.

4. CONCLUSIONS

Design knowledge management by information mining was described. For this purpose, the presentation revealed a development in knowledge acquisition and modeling using *machine learning* methods. In a complex information environment, for specific tasks such as a specialized design task (underground building design, in the present case) with a rich set of design information, a systematic elicitation and account of information in the form of a relational model is necessary. By means of this, consistent design is achieved which matches predefined design criteria. With an actual case study, the research indicated the effectiveness of the present information mining and knowledge management approach in building design. Eventually, this implies the effectiveness of the method as an approach for knowledge-enhanced design. Since the structured information presents a model, which is based on design information and implemented by machine learning, the knowledge base so formed becomes free from subjective decision-makings.

5. REFERENCES

Baarda, D.B. and Goede, de M.P.M. , 1997, *Basisboek Methoden en Technieken,* Educatieve Partners Nederland, BV, Houten

Bezdek C. (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York

Broomhead, D.S. and Lowe, D., 1988, "Multivariable Function Interpolation and Adaptive Networks", *Complex Systems, 21, p.321-355*

Dalen, van J. and Leede, de E. , 2000, *Statistisch Onderzoek met SPSS for Windows*, Uitgeverij Lemma BV, Utrecht

Duda R.O., Peter E.H. and David G. Stork, 2001, *Pattern Classification*, Wiley Interscience, New York

Durmisevic, S., Ciftcioglu, Ö. and Sariyildiz, S., 2001, "Knowledge Modeling by Artificial Intelligence for Underground Space Environment", *Conference Proceedings EuropIA8, 25-27 April 2001, Delft, The Netherlands*

Elenayar S. and Shin Y.C., 1994, "Radial Basis Function Neural Network for Approximation and Estimation of Nonlinear Stochastic Dynamic Systems", *IEEE Trans. Neural Networks, Vol.5, pp.594-603*

Chen S, C.F.N. Cowan and Grant, P.M., 1991, "Orthogonal Least Squares Algorithm for Radial Basis Function Networks*", IEEE Trans. Neural Networks, Vol.2, No.2, March*

Chen M.S. and Manry M.T., 1993," Conventional Modeling of the Multilayer Perceptron Using Polynomial Basis Functions", *IEEE Trans. Neural Networks, Vol.4, pp.164-166*

Ciftcioglu, Ö., Durmisevic, S. and Sariyildiz, S., 2001, "Multi-Resolutional Knowledge Representation", *Conference Proceedings of EuropIA8, 25-27 April 2001, Delft, The Netherlands*

Han J. and Kamber M., 2001*, Data Mining: concepts and techniques*, Morgan Kaufmann, San Francisco

Haykin S., 2000, *Neural Networks:a comprehensive foundation*, Prentice-Hall, Upper Saddle River

Hunt K.J., Haas R. and Murray-Smith R., 1998*,* "Extending the Functional Equivalence of Radial Basis Function Networks and Fuzzy Inference Systems", *IEEE Trans. Neural Networks, Vol.7, No.3, May*

Jang J.S.R. and Sun C.T., 1993, "Functional Equivalence Between Radial Basis Function Networks and Fuzzy Inference Systems", *IEEE Trans. Neural Networks, Vol.4, No 1,*

Krzysztof J.C, Pedrycz W. and Swiniarski R.W., 1998, *Data Mining Methods for Knowledge Discovery*, Kluwer Academic, Boston

Leonard J.A., Kramer M.A., and Ungar L.H., 1992. "Using Radial Basis Functions to Approximate a Function and Its Bounds", *IEEE Trans. Neural networks, Vol.3*

McGrew, C. J. & Monroe C. B. 1993, *An Introduction to Statistical Problem Solving in Geography*, Wm. C. Brown Communications, Inc., Dubuque

Moody J. and Darken C., 1988, "Fast Learning with Localized Receptive Fields", in Proc. Connectionist Models Summer School, D. Tourezky, G. Hinton, and T.Sejnowski (eds.), Carnegie Mellon University, Morgan Kaufmann Publishers

Park J. and Sandberg I.W., 1991. "Universal Approximation Using Radial Basis Function Network", *Neural Computation, Vol.3, pp.246-257*

Saltelli A., Chan K. and Scott E. M., (Eds.), 2000. *Sensitivity Analysis*, Chichester:Wiley

Sowa, J. F., 1985, *Conceptual Structures*. Reading, MA: Addison-Wesley

Tanaka K. ,1997, *An Introduction to Fuzzy Logic for Practical Applications*, Springer, Berlin